ISSN: 0828-2412

A Comprehensive Approach To Public Policy Evaluation: The Implementation-Outcome Connection

Erasmus U. Morah

DP #21
July 1990

UBC PLANNING PAPERS
Discussion Paper #21

School of Community and Regional Planning
University of British Columbia
6333 Memorial Road
Vancouver, B.C. V6T 1W5
Canada

Dr. E. Morah is associated with the U.B.C. School of Community and Regional Planning, where he received his doctorate in Development Planning and Administration. He holds a Master's degree in Public Policy and Public Administration from Concordia University, Montreal.

91-00621/

TABLE OF CONTENTS

1. INTRODUCTION 1

2. BASIC INFORMATION ON POLICY EVALUATION 4

   2.1 Defining Public Policy 4
   2.2 The Emergence of Evaluation Research 6
   2.3 The Role of Social Science 7

3. STAGES IN OUTCOME EVALUATION 9

   3.1 Identification of Goals and Objectives 9
   3.2 Construction of a Causal Impact 11
   3.3 Choice of Indicators of Effectiveness 14
   3.4 Development of an Appropriate Research Design 15
   3.5 Data Collection 23
   3.6 Data Analysis and Interpretation 26

4. THREATS TO THE OUTCOME EVALUATION APPROACH: THE ARGUMENT FOR IMPLEMENTATION EVALUATION 32

   4.1 The Threat of Accuracy 32
   4.2 The Threat of Non-Utilization 36

5. TYPES OF IMPLEMENTATION EVALUATION AND THEIR METHODS OF CONDUCT 39

   5.1 Program Documentation 39
   5.2 Performance Monitoring 41
   5.3 Implementation Assessment 42
   5.4 Process Evaluation 44

6. SUMMARY AND CONCLUSION 45

   6.1 Summary 45
   6.2 Conclusion 46

1. INTRODUCTION

This is a "how-to" paper on program evaluation as a core method of analysis in the field of public policy. The method "embraces" a variety of approaches which are said to be distinctive insofar as they use social science techniques to undertake the analysis (Rossi and Wright 1977, p. 171). This paper emphasizes only two of these approaches -- (1) outcome evaluation focusing on quasi-experimental designs and (2) process evaluation focusing on implementation.[1]

The choice of these two approaches stems from two reasons. The first is that "policy is a process as well as an outcome. It is used to refer to the process of decision-making and also to the product of that decision" (Wildavsky 1979). The second is that other *ex ante* evaluation techniques such as cost benefit, social or environmental impact assessment are largely beside the point. Politics, not evaluation results, determine which policies are to be adopted, continued or terminated. "Evaluations can influence fine-grain program decisions, and that's about it. For any of the big policy decisions, research is marginal. These yea/nay decisions are made on the basis of political tea leaves" (Datta 1976, p. 174).

Nevertheless, once policies are indulged for whatever reasons, there becomes a more genuine need on the part of all those concerned to ascertain whether the policies are achieving the desired results and, if not, how they can be made to do so. Used here, therefore, evaluation research refers to "judgments about the effectiveness of past actions, not appraisal of possible future alternative strategies" (Smith 1981, p. 225). It involves "the use of scientific methods to

1. These two approaches are sometimes referred to as summative and formative evaluation, respectively. The former is said to "aim at determining program results and effects, especially for the purposes of making major decisions about program continuation, expansion, reduction, and funding;" whereas the latter aims at "assessing information for program improvement, modification, and management" (ERS 1980, p. 3-4).

measure the implementation and outcomes of programs for decision-making purposes" (Rutman 1984, p. 10).

In discussing the two approaches of interest, it will be argued that an outcome evaluation, without a preliminary evaluation of implementation, is limited both in utility and explanation. As noted by analysts, "it is more likely that impact evaluations not guided by information about program implementation will be poorly designed" (Sorg 1981, p. 149) and "ask the wrong questions" (Williams and Elmore 1976, p. xiii). Evidently, there can be no way of knowing ahead of the evaluation, how much of the program had been implemented, what brand of it was implemented or how well it was implemented. By itself, an outcome evaluation can only produce findings indicating the extent of policy success or failure, but not the reasons for that success or failure. Implementation evaluation, it will be maintained, can be used to "patch up" these shortcomings (Weiss 1972, p. 72), bearing in mind that an ideal evaluation includes both outcome and process. For instance, policy brand or goals often become distinct during actual implementation, and information gathered during this time provide possible explanations for policy success or failures, as well as an invaluable input for decisions about how to improve policy.

It is, therefore, the aim of this paper to examine quasi-experimental methods for conducting outcome evaluation, the shortcomings inherent in the approach and the role of implementation evaluation in responding to these shortcomings. The learning objective for the paper is to be able to formulate research designs that are not only capable of determining program outcomes (in the absence of true experimental designs), but also the explanations behind the outcomes. For concreteness and clarity, the paper draws on the

Ontario seat belt legislation for its examples and illustrations. This legislation has been studied using a similar method (Musto 1985) and readily lends itself to the approaches being elaborated.

In addition to providing some basic information on public policy evaluation, the first half of the paper presents a systematic approach to research on outcome evaluation. The other half depicts the shortcomings of this approach including ways of controlling for them using implementation evaluation techniques.

2. BASIC INFORMATION

2.1 Defining Public Policy

The objective here is not to define public policy proper, but rather to highlight the important dimensions of the term, as well as debates in which it is shrouded. To this end, the approach by Chandler and Chandler (1979, p. 2) appears useful. The term "public policy" is first split in two to examine the constituent concepts separately, and then the disparate meanings of the concepts are re-synthesized once again. The first question of importance -- what do we mean by "public" -- that derives from the approach is readily answerable. While there can be no denying that private or non-public actions by individuals and corporations affect many people collectively (Nadel 1975), the consensus among most writers is that the "public" in public policy refers to actions taken by government or by those who occupy official positions in it (Adie and Thomas 1982, p. 89).

The core controversy over the meaning of public policy surrounds -- what do we mean by "policy?" By it, do we mean (1) a sector of activity; (2) a general statement of purpose or desired state of affairs; (3) a specific program for solving a problem; (4) a decision of government; (5) an output; (6) an outcome; (7) a theory, model or paradigm; and (8) a process (Hogwood and Gunn 1984)? There are no simple answers as the term can be related to each of these meanings.

When it is not the intention of government to ignore problems, policies are usually stated in sectors of activities (e.g., economic or energy sectors) and generally involve statements of goals or desired state of affairs in these sectors (e.g., full employment or self sufficiency). The realization of these desired conditions then depends on the development of action programs

and projects which, from the temporal sequence of events, implies a theory or causation. The implication is that, if certain action programs are implemented, the desired state of affairs will result. But, however, given that desired conditions are rarely what they seem nor materialize as desired, policy also implies what is actually delivered (output) and the outcome of this irrespective of desires. Also, the fact that all the aforementioned activities occur in a particular manner or stages of process implies yet another meaning, as does the fact that the process frequently culminates in a decision.

While policy probably implies these meanings and more, the following clear statements can nevertheless be made. First, policy is not a sector of activity although it is often stated in one. Second, it is not a program as more than one program can be (and frequently is) developed in response to the same desired condition.[2] Third, it is not a decision as it often involves several (smaller) decisions made over time. And, fourth, it is not a theory but necessarily contains a tactical component (i.e., programs) which implies causation. Policy involves more a general statement of purpose or intentions but, because intentions have unforeseeable consequences and may be to change or perpetuate existing conditions, it also involves (in)actions and the (un)intended outcomes of these.

Thus, in uniting the meaning of "public" with that of "policy," we reach the conclusion that "public policy" is "what governments choose to do or not to do" (outputs) (Dye 1978), why they do it (intentions), how they do it (process) and what difference it makes (outcomes).

---

2. Note, however, that policy is a function of programs and is dependent upon their outcomes. The evaluation of policy almost necessarily involves investigation and analysis of concrete action programs designed to achieve broader policy aims. This allows for the two terms to be used inter-changeably as is done throughout this paper.

2.2 The Emergence of Evaluation Research

The practice of evaluation research, as a distinct social science endeavor aimed at rationalizing policies in the public domain, is scarcely three decades old. The need for such a rigorous method of determining the "effectiveness" of public programs arose in the U.S. during the 1960's, when it became evident that the Johnson Administration's War on Poverty, with its awesome expenditures and assortment of programs, was not bringing about desired effects. The events of this era had brought with it two hard lessons, the first being that there is not enough money to do all the things that require doing, and the second that, even if there were enough money, it takes more than money to solve complex human and social problems.

According to analysts (Patton 1978, p. 16), evaluation research was born out of this era and these two lessons, as it became necessary for there to be both a basis for deciding which things are to be done, as well as a deeper understanding of those social problems which are the objects of policy. Much early efforts (as with current ones) in this regard were carried out in the fields of education, health and legal reform as these areas were thought to be arenas in the War on Poverty. Consequently, techniques comprising evaluation research are more developed and adapted for policies from these fields.

As might have been expected, the emergence of evaluation research in Canada came as a result of "spillover" effects from the U.S.. But, unlike in the latter, its emergence in Canada was more a political response to widespread demand by the Canadian public for more "accountability" over the management of the public purse (Jordan and Sutherland 1979). During the early 1970's, accountability had become a "buzz" word in Canadian politics, necessitating the federal government in 1976 to respond by creating the Office of the Comptroller General within the Treasury Board. The broad function of this office was to tighten-up financial management, and related administrative practices and control throughout

the federal bureaucracy. In particular, the office was to ensure that the three E's in government (effectiveness, economy and efficiency) were measured and maintained by those government agencies that execute programs.

2.3 The Role of Social Science

Right from its emergence, the basic assumption of evaluation research was that social science knowledge and methods could help towards making better policy decisions (Rein 1976). As mentioned earlier, the history of the Great Society programs had demonstrated that part of the problem in public policy was not understanding the problem in the first place. For instance, if the causes and complexities of poverty were never fully understood, how then could poverty have been effectively stamped out? It is basically in this regard that it was felt that social science could come to the rescue and provide answers. The expectation was that techniques from this area could be enlisted to provide the rigor necessary to satisfy the new requirement for tougher evaluation standards which, in turn, would make it possible for decision makers to "get the biggest bang for the public's dollar."

As Weiss (1977, p. 4) explains, "there was much hoopla about the rationality that social science would bring to the untidy world of government." First, it would generate hard data for planning in general. Second, it would provide evidence of resources but mainly of need, so that decision makers would be able to priorize among competing alternatives and, consequently, choose the morally right course of action. Third, it would provide cause-and-effect theories, so that planners would know which variables to recommend for manipulation in order to effect the desired outcomes. Fourth, it would bring to the assessment of alternatives dispassionate reasoning and a knowledge of relative costs and benefits, so that choices could be made between options on the basis which option had the highest payoff.

And, fifth, once policies were in operation, it would provide an objective evaluation of their effectiveness, so that necessary modifications could be made to improve performance. The underlying assumptions were that fact (means) and value (ends) could be separated procedurally, thereby making dispassionate reasoning and optimal solutions possible; that causality could be effectively modelled along with the external conditions under which the causality might apply; and that decision makers would mainly be interested in rational or optimal paths to solutions.[3]

3. For an iconoclastic discussion of these assumptions and what is involved in the assertion that social science can assist in rationalizing policies, consult Rein's (1976) book on ***Social Science and Public Policy***.

3. STAGES IN OUTCOME EVALUATION

3.1 Identification of Goals and Objectives

Although it is a truism that policy goals are by nature vague, diffuse, ambiguous and contradictory, it is equally a truism that a "program which does not have clearly specified goals cannot be evaluated without specifying some measurable goals" (Rossi 1972, p. 18). "If evaluators agree on anything, it is that program objectives.... (are) an extremely important consideration when an evaluation study is constructed" (Worthen and Sanders 1973, p. 231), for without the basis that the goals provide, an evaluator will flounder and not be able to ask the relevant and insightful questions about the program. Moreover, the core outcome evaluation question remains -- "To what extent is the program succeeding in reaching its goals" (Weiss 1972, p. 26)?

As such, "any evaluation -- whether planned in advance, forming part of an ongoing operation or constituting an ex-post review of project implementation -- must therefore begin with a thorough analysis of the objectives of the project or program concerned" (Grabe 1983, p. 27).[4] This often entails (re)stating the program goals in such a way that numerical expressions can be found which accurately capture the essence of the goals. While this does not mean an insistence on quantifiable goals, it means being able to get beyond the formal goals of the program to identify its more concrete and operational objectives/targets. The questions forming part of the goal analysis stage include:

---

4. It should be noted that there has been a daring move for "goal-free evaluation" led by Scriven (1971, p. 1-2), who not only finds the consideration of goals in evaluation unnecessary but possibly contaminating. He argues that "the less the external evaluator hears about the goals of the project, the less tunnel vision will develop, the more attention will be paid to looking for actual effects (rather than checking on alleged effects)."

(1) What are the general development policies under which the program was designed?

(2) Does the program constitute a single effort towards attaining these objectives or does it form part of a wider program consisting of several program elements?

(3) To what target population is the program activities directed?

(4) What are the immediate objectives pursued under the program?

(5) What time-limits have been set for attaining each objective of the program (Grabe 1983, p. 27)?

To determine the intent of policy directives on these, Nakamura and Smallwood (1980, p. 74) suggest using cues from the following:

(1) The "legal intent" of the policy as written.

(2) The "legislative history" of the policy.

(3) The "understanding" of a policy held by interest groups and other attentive publics.

(4) The goals "developed" by participants during policy implementation.

(5) The "translation" of a policy's goals from vague political language to more precise terms by evaluators.

(6) The policy goals most consistent with the research requirements of evaluators.

(7) Current policy maker understandings of policy goals.

In terms of the Ontario seat belt legislation, the objective of the policy can be established using cues 1 and 2. The "legal intent" of the policy as written (in Ontario's Bill 27 which came into effect on January 1, 1976) was to make wearing of seat belt mandatory for all motorist and passenger population.[5] The "legislative history" of the policy related to the escalating deaths of Ontarians due to motor vehicle collisions, as evidenced by a 20 per cent increase between 1970 and 1973. It was, therefore, the goal of the legislation to get motorists to

5. Exceptions were made for: (1) driving in reverse; (2) persons under sixteen years of age; (3) persons who, because of health or physical size, were unable to wear a seat belt assembly; and (4) persons who were actually engaged in work which require them to alight from and re-enter a motor vehicle at frequent intervals and who, while engaged in such work, do not drive at a speed more than 25 miles per hour.

buckle up in the short range, reduce fatality rates in the medium range, and possibly change attitude towards automobile safety in the long range. Whether or not the legislation represented a single effort towards attaining this objective is questionable. The government of Ontario, prior to introducing the legislation, had launched a $650,000 information program to persuade motorists to wear seat belt voluntarily. It also simultaneous with the seat belt legislation, reduced the speed limit by 10 miles per hour and "cracked down on drunk drivers."

3.2 Construction of a Causal Impact

After having delineated the goals and objectives of a policy, the next step is to build an impact model by identifying the chain of reasoning and assumptions that led to the adoption of the course of action in question. Every policy implies a theory or causal relationship, observed Pressman and Wildavsky (1973, p. xiv). "The question never is whether a theory is there (it always is).... A promise underlies policy: if the actions we recommend are undertaken, good (intended) consequences rather than bad (unintended) ones actually occur" (Wildavsky 1979, p. 35). In other words, policy is an "if-then" statement in the sense of -- "if X is done, then Y will result."

Constructing an impact model for the purposes of evaluation research, therefore, requires an abstraction of this inherent theory of policy. Doing this not only aids the researcher in asking the relevant cause-and-effect relationships questions, but it also enables him to determine early in the evaluation whether or not a program's means for achieving goals are appropriate. As analysts are beginning to learn, it is not unusual that many policies fail, because their means and ends are not connected based on a valid theory of connection (Nakamura and Smallwood 1980, p. 75).

Nachmias (1979, p. 9) identifies three steps involved in the construction of an impact model. The first is to outline the variables that are relevant to the problems under study. The second is to explicate the significant relationship among the variables. And the third is to formulate propositions regarding the nature of these relationships. These propositions may be in the form of one or more testable hypotheses "concerning the changes that the policy is expected to produce, and how the changes will affect the behavior or the condition intended to be modified" (Nachmias, 1979, p. 10).

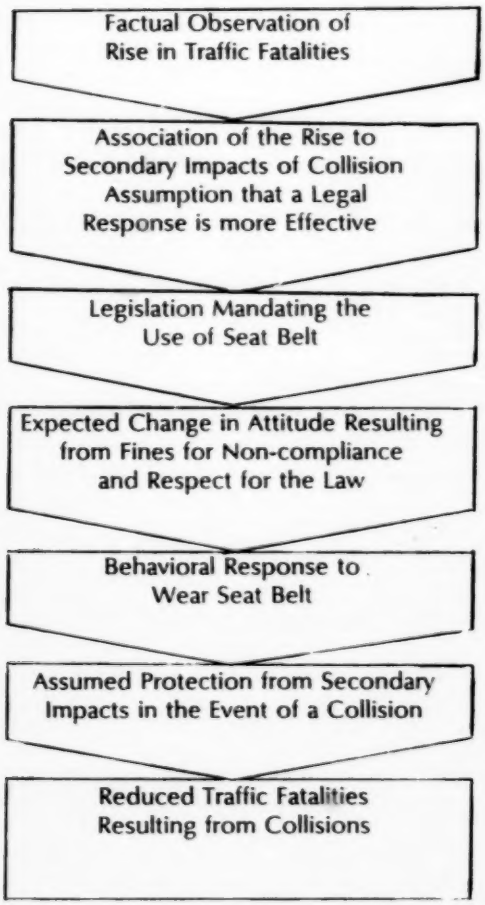
The accompanying figure presents a possible impact model for the Ontario seat belt legislation. It commences with the factual observation that traffic fatality in the province was on the increase, 20 per cent between 1970 and 1973. This observation was then followed by two things: (1) the authority's association of the increase to secondary impacts of collisions resulting from being thrown around or out of the vehicle;[6] and (2) the assumption that non-legal responses (e.g., public education) to persuade motorists to voluntarily wear seat belt would work less well than a legal response based on punishment for non-compliance and obedience out of respect for the law. From both this association and assumption emerged the seat belt legislation which made their use compulsory for all motorists. The lap belt was designed to prevent the motorist from being ejected from the vehicle, while the shoulder belt restrains his head/chest from impinging on the steering wheel or dashboard. The remainder of the causal reasoning is self-explanatory.

---

6. Note that this association automatically excludes other potential contributors such as speeding, drunk driving, reckless driving, an increase in the number of motorists, etc..

FIGURE 1

A Possible Causal Impact Model for the Ontario Seat Belt Legislation


Factual Observation of
Rise in Traffic Fatalities
Association of the Rise to
Secondary Impacts of Collision
Assumption that a Legal
Response is more Effective
Legislation Mandating the
Use of Seat Belt
Expected Change in Attitude Resulting
from Fines for Non-compliance
and Respect for the Law
Behavioral Response to
Wear Seat Belt
Assumed Protection from Secondary
Impacts in the Event of a Collision
Reduced Traffic Fatalities
Resulting from Collisions

3.3 Choice of Indicators of Effectiveness

Once policy goals and objective have been stated, and a causal impact model constructed, the proceeding action is to identify quantifiable indicators or measures which can sensibly be related to the policy. This is the question of what to look upon as representing policy success, and then developing a yardstick for measurement based on this. The special importance of this stage of the evaluation is that the kinds of goal measures used will tend to determine the degree to which the policy's performance is judged satisfactory. Yet the choice of indicators of goal accomplishment is a profoundly arbitrary process, making it difficult to develop agreement on which operational indices validly represent established goals and objectives.

No strict guidelines exist for selecting indices of a policy's success, except that whatever indices chosen should:

(1) follow from the goals and objectives of the policy;

(2) be consistent with the causal impact model;

(3) have high degree of validity and reliability;[7]

(4) take into consideration perceived impacts of policies by relevant attentive publics (Mazmanian and Sabatier 1979); and

(5) be structured along the three dimensions of policy -- outputs, outcomes and impacts -- as suggested by Levy, Meltsner and Wildavsky (1974, p. 4-8).

However, in order to achieve a consensus on any given set of measures of effectiveness, it is recommended that an approach similar to the one suggested for establishing policy goals and objectives be used.

---

7. In the context of evaluation research, Grabe (1983, p. 32) defines validity to mean the measurement of a specific characteristic, state or process that is, in fact, the object of measurement. Reliability simply means the ability of the measurements to yield consistently stable results.

Indicators relating to output remain the most widely used, as output is the easiest of the three dimensions of policy to measure quantitatively. It merely requires an enumeration of the (most immediate) units of service actually delivered. In the case of the seat belt legislation, output performance indicators would include the legislation itself, its promulgation (i.e., the number of people who knew about the law) and enforcement (e.g., the number of fines handed out for driving without seat belt). Outcome indicators, which deal with cause-and-effect relationships of the policy in human terms, would include the percentage of motorists who were buckling up since the legislation, and the consequent percentage reduction in the number of traffic fatalities resulting from collisions. The long(er) term consequences of the policy or its impacts, may be indicated by change in attitudes towards automobile safety or safety in general.

3.4 Development of an Appropriate Research Design

When policy goals have been established, a causal impact model constructed and indicators of effectiveness selected, it becomes necessary to define an overall research design to guide the actual process of collecting, measuring, analyzing and interpreting data about the program. Three distinct broad categories of designs can be identified: (1) non-experimental designs, (2) experimental designs and (3) quasi-experimental designs.

3.4.1 Non-Experimental Designs. Tien (1979, p. 500) identifies three types of such designs, the most well known type being the "case study" design diagrammed below, where X stands for a program, treatment or intervention, and O for an observation.

X O

The design typically involves a post-mortem analysis of a program "with the objective of highlighting the salient points for the purpose of learning from the experience." Another less well known but frequently used type is the "survey study" design which, much like a term paper design, involves "a general analysis of similar data from several sources (representing, for example, a topic area) for the purpose of drawing conclusions about the combined sources (or topic area)." Finally, there is the "cohort study" type design which focuses on "recording the performance of a group of individuals or subjects over time" (Tien 1979, p. 500).

The distinguishing characteristic of these design types is that they do not make use of any control; a fact which automatically makes whatever findings they may yield unscientific or suspect, to say the least. As Ross and Wright (1977, p. 17) put it, even though such studies "can be valuable simply because some information is always better than even the best uninformed guesses, the absence of control makes it impossible to discern whether or not the observations reflect the effects of treatment or of any of a host of conceivable alternative competing factors." That is why, despite the popularity of these design types arising from the fact that they are fast and easy to use, the clear suggestion is that "they should not be used if other alternative (i.e., experiment and/or quasi-experiment) designs are possible" (Tien 1979, p. 501).[8]

3.4.2 Experimental Designs. These are the polar opposites of the non-experimental designs, consisting of four common features: (1) comparison, (2) manipulation, (3) control and (4)

---

8. With respect to the "The One-Shot Case Study," it should be noted that, while its obvious deficiencies are the lack of pretest observations and a control group, Cook and Campbell (1979, p. 96) point out that case studies are often carried out in settings where many variables are measured at the pretest; where contextual knowledge is already rich; and where intelligent presumptions can be made about what the target group would have looked like without the treatment. These three factors, they argue, "can often serve the same role that pretest measures and control groups do."

generalizability (Nachmias 1979, p. 23). The core of the designs, however, is the use of controls and "the random assignment of units (persons, classes, cities, and so on) to two groups -- an 'experimental' and a 'control' group" (Rossi and Wright 1977, p. 14).

The random sorting of subjects into groups guarantees more than any other technique that subjects will not vary on the average (i.e., more than could be expected by "pure chance") with respect to any particular characteristics prior to being exposed to the experimental conditions, program or treatment. If both the experiment and control groups are not brought under identical conditions, except for their differential exposure to the treatment to influence observation, then grounds for inferring causality do not exist.[9] In most normal circumstances, the experimental group is administered the treatment and the control group a placebo or nothing. And, depending on whether the experiment is double or single blind, the experiment administrators may or may not know which group receives what treatment.

Hence, in its ideal form, a true experimental design involves administering a program or treatment to one of two equivalent groups, and then observing both groups over time for variations. Whatever subsequent differences that emerge between the two groups are held to be the effects of the treatment. Or the only plausible explanation of the differences outside of chance factors in the randomization that may have made the experimental and control groups non-equivalent. This latter scenario is ruled out through statistical inferences and assumptions.

---

9. Needless to point out, the notion of causality differs from that of correlation or statistical association. That a change in X is followed by a change in Y does not mean that X causes Y or vice versa. Such might be the suggestion but, to establish a cause-and-effect relation, it has to be ruled out that the change in Y is not owed to other variables outside of X such as chance, bias or error. One caveat about this distinction between causality and correlation, however, is that the former belongs to the theoretical domain as cause-and-effect relationships cannot be observed, while correlations can.

"There is almost universal agreement among evaluation researchers that the randomized controlled experiment is the ideal model for evaluating the effectiveness of a public policy" (Rossi and Wright 1977, p. 13). This should hardly be surprising given that the approach is the backbone of science, and provides a far more superior inferential power than any other known design. However, it is not always feasible to use randomized or true experimental designs for outcome evaluations. In fact, with the exception of "social experimentation" in which programs are tested and evaluated before being implemented, most public programs cannot be evaluated using true experimental designs. The fact that most public programs are universally applied, makes it nearly impossible to control the study population, let alone the environment. Furthermore, most evaluations conducted are on programs that are already in place, which makes it unethical and, probably, legally impossible to withhold treatment to a would-be control group who may merit or need the treatment more.

Additional to these impediments is the problem that arises from the political imperatives of policy. "Well conducted field experiments ordinarily take a number of years to conduct and still additional years to analyze" (Ross and Wright 1977, p. 15). Firstly, this means that the approach can only lend itself when there is both sufficient funds to cover their unusually high cost and no urgent need for the program, which is probably never. Secondly, it also means the exacerbation of the general "time-lag" problem in which the experimental setting, reality or problem may change considerably before a consequential policy from the experiment is adopted.

Furthermore, the very fact that experimental programs are likely to be better-conducted than real world programs also poses a problem. It would follow that the effects observed are apt to be of some idealized version of the program which may not hold in the "real" world. Finally, there is the methodological pitfall arising from the fact that "it is often very difficult to

maintain an experimental design over time without suffering severe mortality or attrition, especially of the control group" (Ross, Wright and Wright 1978, p. 180).

3.4.3 Quasi-Experimental Designs.[10] For the aforementioned reasons, particularly the impossibility or impracticality of randomization and the political imperatives of policy, attention in evaluation research has focused on quasi- or almost-experimental designs. The designs are so-called because they embody treatments, outcome measures and experimental groups as do true experiments. Their crucial point of departure is that they do not make use of randomization to produce equivalent experimental and control groups. Instead, they make use of non-random or non-equivalent groups that differ from each other initially, in addition to their differential exposure to the treatment. As such, "the task confronting persons who try to interpret the results from quasi-experiments is basically one of separating the effects of a treatment from those due to the initial non-comparability between the average units in each treatment group" (Cook and Campbell 1979, p. 6).

It should be noted, however, that it is not necessary as much as it is preferable that all quasi-experimental designs utilize one form of non-random control or another. In fact, quasi-experimental designs such as "the interrupted time series" designs are considered especially useful in cases when no comparison group can be found. In "cross-sectional" studies, for another instance, no attempt is made whatsoever to use a control. Competing explanations for observed effects are controlled statistically using multivariate analyses which hold the competing factors constant. Thus it is that, in addition to relying on non-random controls for distinctiveness, quasi-experimental designs also rely on the identification of threats (i.e.,

---

10. According to the Rossi and Wright (1977, p. 13), if there is a Bible for this method, "the Scriptures have been written by Campbell and Stanley (1966), along with a revised version by Cook and Campbell (1979)."

alternative explanations for observed effects) and ways of controlling for them. Five quasi-experimental designs that dominate both the literature and practice are described below.[11]

The One-Group Pretest-Posttest Design is one of the most widely used designs in evaluation and is symbolized below. ("O" stands for an observation; "X" stands for a treatment; and "1" through *"n"* refer to the sequential order of recording observations.)

$$O_1 \quad X \quad O_2$$

This design requires observations of the target population, administration of a treatment or program and observations of changes in posttreatment values of the target population to estimate effects. If the experimental treatment had any effect, there will be discrepancies in observations before and after the intervention. Although, depending on the nature of the problem being treated and the magnitude of the intervention, the discrepancies may not be readily discernible, as posttreatment effects may be immediate, delayed, continuous or discontinuous.

Despite the fact that the "the one-group pretest-posttest" is the most widely used by government evaluators who are often faced with budgetary and time constraints, the design has little scientific relevance. It uses no form of control to eliminate alternate causal interpretations and, as such, is only slightly preferable to the non-experimental designs which were said to hold no scientific value whatsoever (Tien 1979, p. 502).

---

11. Excluded from discussion are those designs that do not lend themselves to the evaluation of programs already in place. Examples are designs with removed or switching replications of treatment. Obviously, the political nature of public programs would not permit for alterations of intervention (goods/services) nor its removal in the form of non-intervention. Public programs are often justified on the basis of need, merit and/or equity.

The Posttest-Only Design with Non-Equivalent Groups is mainly used when evaluators are unable to work out their research design until after the treatment had begun. Being *ext post facto* as such, the design (diagrammed below) relies on posttreatment measurement and a non-equivalent comparison group.

X O

O

An example might compare the traffic fatality rate in Ontario after the province's seat belt legislation with the rate for another similar province (e.g., Quebec) which did not have an equivalent legislation. The obvious problem with such a comparison would be that the two provinces are likely to differ initially with respect to traffic fatality rates. No pretreatment observations on the two provinces would be another point of difficulty but, as Cook and Campbell (1979, p. 98) point out, this should not be a significant concern "for archival records can often be used to establish what the pretest scores of the various experimental units were." Of course, such measures can only be correlational with posttreatment measures, as instruments of measure are likely to differ for the two stages.

The One-Group Pretest-Posttest Design with Non-Equivalent Groups subsumes the first design mentioned and is recommended as the preferable of the two. For research in social science, this is "perhaps the most frequently used design.... (it is) recommended in situation where nothing better is available" (Cook and Campbell 1979, p. 104). Represented below, the design controls for more threats to validity than the two preceding designs.

O_1 X O_2

O_1 O_2

It does not, however, control for problems related to non-randomization, regression artifacts[12] and trends (seasonal/cyclical or in other ways) which might otherwise be masquerading as a treatment effect. This is what makes this next design worthy of mention.

The Simple Interrupted Time Series Design is diagrammed below. The design requires no external control, but its requirement for multiple observations (at least three observations in general) before and after an intervention, qualifies it as its own control; a control which, in all respects, can be considered equivalent.

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad O_5 \quad X \quad O_6 \quad O_7 \quad O_8 \quad O_9 \quad O_{10}$$

Although, without an additional control, the design may not be preferable to the preceding one. According to Campbell and Stanley (1966, p. 39), the design fails to control for those threats arising from changes in observations due to effects-causing-events other than the treatment which may occur during the pretreatment and posttreatment period.

An example of the (dis)advantages of the design are the public statements by the then Ontario Transport Minister, James Snow, regarding the effects of the seat belt legislation. A ***Toronto Star*** (1977) article based on Snow's statements, explained that there were 223 road deaths in the first four months of 1977, whereas in the same period in 1976, there were 240 fatalities. The article went further to state that, in 1975, before the seat belt were made compulsory, there were 415 road deaths in the same four month period. The advantage in this stems from Snow's mention of three separate observations, rather than a simple before and after observations. The disadvantage, however, is in the fact that there is a number of

12. This is a threat that results when groups or cases are chosen for treatment (as is often the case) because their dependent measures have extreme values. The hypothesis is that these extreme values are not real and would, therefore, tend to regress to the mean of the sample without any intervention. Recall that the Ontario seat belt legislation was enacted following a sharp increase (20 per cent in three years) in traffic fatalities.

extraneous events which may preclude Snow from attributing the reduced traffic fatalities to the effects of the legislation. For example, the weather conditions following the year of the legislation, if unusually dry, may have resulted in fewer accidents that would normally be due to rain, snow or fog. Furthermore, given that traffic fatality rates fluctuate even when there is no intervention, it could be argued that this normal instability may have been responsible for the decline in fatalities. These threats are controlled for in the next design, which is perhaps the strongest of the many possible quasi-experimental designs. In best experimental conditions (e.g., when data, time and money are not a problem), this is the recommended design.

The Interrupted Time Series Design with Non-equivalent Groups can be demonstrated as:

$$O_1 \; O_2 \; O_3 \; O_4 \; O_5 \; X \; O_6 \; O_7 \; O_8 \; O_9 \; O_{10}$$
$$O_1 \; O_2 \; O_3 \; O_4 \; O_5 \qquad O_6 \; O_7 \; O_8 \; O_9 \; O_{10}$$

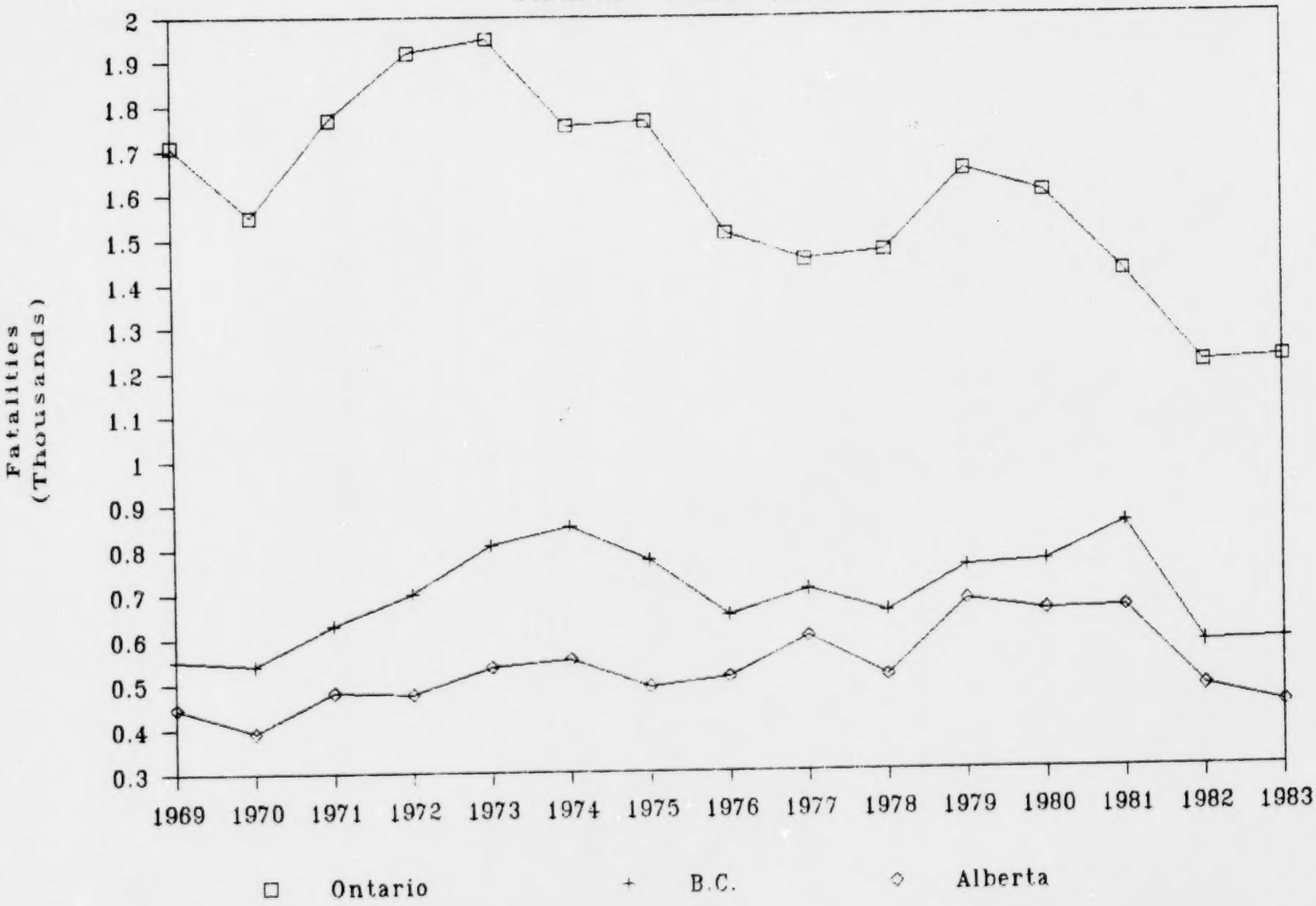
It subsumes the preceding design by adding to it yet another control, although this added control is not equivalent. An application of this design to the Ontario seat belt legislation can be found in the accompanying graph. The design controls for all the threats to internal validity.

3.5 Data Collection

Once the goals and objectives are understood, a causal impact model constructed, measures of effectiveness established and the appropriate overall research design selected, the researcher is then confronted with the task of having to identify where data for the evaluation can be obtained from. While recognizing that "there is no one source nor one method that

Fatal motor vehicle accident victims

Statistics Canada (1984)



exclusively suits evaluation research.... (and that) the kinds of data to be collected are governed by the nature of the policy to be evaluated, the kinds of variables included in the impact model, and the research design" (Nachmias 1979, p. 17-8), the following are the more commonly used sources and methods:

> 1. *Self reports* -- you can ask the participants or target population about themselves and their experiences.
>
> 2. *Reports from others* -- you can ask others who know the participants about them and their experiences.
>
> 3. *Direct observations* -- you can observe the participants.
>
> 4. *Output measures* -- you can check program and operation records for the amount of services actually delivered.
>
> 5. *Performance outcomes* -- you can test participants for the presence of effects (Dimock 1979, p. 8).
>
> 6. *Social indicators and official statistics* -- you can check relevant quantitative indices of societal conditions and trends available on a common accounting basis (Nachmias 1979, p. 80-4).

Any one of these sources and methods could be used to gather information about the effects of the Ontario seat belt law including: (1) archival data from Statistics Canada on "Causes of Death;" (2) survey research establishing the number of motorists and passengers buckling up as a result of the policy; (3) output data from program operation records on the enforcement of the law (e.g., the number of fines handed out for failure to wear seat belt; and (4) random observations of seat belt wearers from busy street intersections.

3.6 Data Analysis and Interpretation

When all the details of the program have been ascertained, questions asked about them and the answers documented, and all the related data gathered, the final step of the evaluation is to analyze and interpret the intelligence gathered in such a way that a valid causal inference can be made about the program effects. While this requires the application of relevant data-analysis techniques, what is required more is knowledge of the various threats to validity[13] common with the chosen research design and ways of controlling for them. Statistical analysis *per se*, as noted by McCain and McCleary (1979, p. 233), "contributes little to ruling out threats to internal, external, and construct validities."

By threats to validity, it is meant alternative explanations or plausible rival hypotheses for observed effects of a treatment or program. Evidently, without the minimization of these threats to validity, an evaluation can at best be expected to have little or no scientific relevance. There are basically five categories of threats to validity. These are discussed below, along with ways of controlling for them.

3.6.1 Internal Validity Threats. These refer to "the extent that the statistical association of an intervention and measured impact can reasonably be considered a causal relationship" (Tien 1979, p. 497). This is the key question, for example, of how can one be certain that observed differences before and after the seat belt law are not a product of "chance" or "extraneous" factors influencing the observations? The answer is that one cannot be certain unless a number of threats are controlled for.

---

13. The classic monograph for much of the discussion on threats to validity, according to Tien (1979, p. 497), is still that of Campbell and Stanley (1966).

Differential selection is a basic threat in all quasi-experimental designs which make use of non-random sampling.[14] The threat posed is that observed effects may be due to initial differences or biases of the two groups rather than to the treatment itself.

The threats posed by selection can be controlled for, first, by analyzing the background characteristics of the selected groups; and second, by analyzing the pretreatment measurements of the selected groups using the elementary analysis of variance (ANOVA) model.[15] According to Reichardt (1979, p. 153) "if the result is not statistically significant, the groups are assumed to be equivalent." However, ANOVA on pretreatment measurements does not resolve threats related to *selection-maturation*, as the group differences, even though initially statistically insignificant, can increase over time in a significant and unrelated way to the treatment.

Cyclical trends are a threat when effects may be due to seasonal tendencies or established pattern of events. For example, how can one be sure that observed differences before and after the seat belt law are not a product of a long-term declining trend or *normal instability* in the time series? From 1973 to the enactment of the law in 1976, traffic fatality rate declined consistently by approximately 7.5 per cent on the average.

Cyclical or seasonal trends (as well as normal instabilities) can be controlled for by having a long time series which would permit for a visual inspection of manifesting patterns. Evidently, the longer the time series the easier it becomes to detect even the slightest trend, pattern or instability.

---

14. It is also a threat in randomized experiments because of *differential loss* of subjects from the test and control groups. This loss, which tends to be higher in the control group, may introduce initial biases.
15. This model is elaborated under "statistical conclusion validity threats."

History is the major threat to internal validity with most single (no external control) time series designs. It is a threat when observed effects might be due to events which occur between the pretreatment and posttreatment period, when these events are not related to the treatment.

It was noted earlier how the following extraneous events may have posed a threat in analyzing the effects of the Ontario law:

> (1) The $650,000 *ex ante* public education campaign by the Ontario government to persuade motorists to wear seat belt voluntarily.
>
> (2) The simultaneous reduction in speed limit by 10 miles per hour.
>
> (3) The crackdown on drunk drivers which, in 1980, was cancelled and replaced by a law which permits the police to issue 12-hour license suspensions to drivers shown to be legally impaired.

The most effective way to control for history threats is to add a non-equivalent no-treatment control group to the design. Another method is to examine the records to see if other effect-causing events could have influenced observations.

Other threats to internal validity include *instrumentation changes* (i.e., non-standardization of measures); *regression artifacts* (i.e., the choice of a target population because its outcome measures contain extreme values); *and temporal maturation* (i.e., changes in observations due to subjects growing older, weaker, stronger, wiser, more experienced and the like). The first threat can be remedied by standardizing the calibration of measuring instruments, while the second and last threats can be controlled for by examining pretreatment data.

3.6.2 External Validity Threats. These refer to "the extent that the causal relationship can be generalized to different population, settings, and times" (Tien 1979, p. 497). The threats

posed arise mainly from the unique setting (i.e., *interaction of setting and treatment*) and time (i.e., *interaction of history and treatment*) in which experiments are conducted.

The former threat can be controlled for by varying settings and analyzing for causal relationship within each. Or by dividing the target population into subgroups (e.g., males, females, age groups, etc) and assessing whether effects hold across the various cohorts. For the latter, either the experiment can be replicated at different times, or a literature review can be conducted to see if prior evidence exist which do not refute the causal relationship of interest.

3.6.3 Construct Validity Threats. These are less of a threat to most experiments or evaluations, except for those that involve multiple measures, interventions and intervenors. They refer to "the fitting of cause operations to their referent constructs"(Musheno 1981, p. 78), or "the extent that the causal relationship can be generalized to different interventions, impact measures, and measurements" (Tien 1979, p. 497). It is not uncommon that what one investigator measures and interprets as a causality (*A* ---> *B*), may be measured and interpreted differently by another investigator (*A* ---> C). This is essentially the problem of "confounding" effects, which derive from the possibility of the investigators' misconstruing what is intended to represent a particular cause and effect construct.

There are three steps to ruling out alternative causal constructs. The first is to test how much the treatment varies with the object of change or what is intended to be altered. The second is to test how much the treatment varies with related but different constructs or objects of change. And the third is to ensure that the treatment is not dominated by impertinent factors, as these can compound to mean more or less of a particular cause and effect construct.

3.6.4 Statistical Conclusion Validity Threats. These threats are concerned with the statistical association of a treatment and an effect (i.e., statistical relationship). A false association is referred to as *type 1 error* and a false non-association as *type 2 error*.

Beyond increasing statistical power by increasing sample size, statistical conclusion validity can be tested and controlled for in three simple ways. The first is through the *ordinary least-squares* (or "best-fitting" line) method to approximate changes in the "level" and "slope" of observations between the pretreatment and posttreatment period. The second test is the *simple analysis of variance (ANOVA)*, whose component elements include:

(1) the grand mean of the posttest scores across all participants;

(2) the treatment effect, which is the average value that the treatment adds to, or subtracts from, the posttest scores in the treatment group; and

(3) the error or residual, which is the distance of each observation plotted on a scattergram from the best-fitting line (Reichardt 1979, p. 151).

And the third test is the extension of the ANOVA model to include pretreatment measures. Known as the *analysis of covariance (ANCOVA)*, the method simply repeats the simple ANOVA on the pretreatment measurements. This way, an adjustment is made for initial differences between the groups.

3.6.5 Conduct Conclusion Validity Threats. These are the newest dimension of threats to validity, involving the extent to which program structure (e.g., complexity or goal clarity), political infeasibility (e.g., institutional resistance) and economic infeasibility (availability of resources) may have affected the overall evaluation effort and result.[16] Trying to imagine what more could have been done or done differently had these threats not been a problem and estimating the probable impact of this on what was actually done, may serve as a control.

16. Conduct validity is distinct from "evaluability assessment," which is an attempt to determine the extent to which a relevant and credible evaluation can be successfully implemented before committing resources to it (Rutman 1984, p. 27).

3.6.6 Approach Validity Threats. These are by no means a set of standard threats, but one proposed by this author to alert that "some approaches (in and of themselves) may not be up to the task for which they are called upon to perform." Earlier, we had defined the purpose of evaluation research as providing accurate and useful information for decision making. Threats to approach validity deals with the extent to which outcome evaluation satisfies this purpose. It is in this regard that many serious questions have been raised about the relevance of the approach. So serious are the questions raised that some critics have urged its abandonment for most types of programs (Weiss and Rein 1970), while others suggest a radical redefinition of its functions (Floden and Weiner 1978). The remaining half of this paper looks closely at this new set of threats and how they can be minimized through implementation evaluation techniques.

4. THREATS TO THE OUTCOME EVALUATION APPROACH: THE ARGUMENT FOR IMPLEMENTATION EVALUATION

As noted by Weiss (1972, p. 43), in a traditional "black box" approach to evaluation, the questions often are:

> Why should the evaluator be concerned with program input? Haven't we noted earlier that his job is to find out whether the program (whatever it is) is achieving its goals? Does it make any difference to his work whether the program is using rote drill, psychoanalysis, or black magic? There are evaluators who are sympathetic to such an approach. They see the program as a "black box," the content of which do not concern them; they are charged with discovering effects.

To the extent that evaluation research is dominated by outcome evaluation, the question of interest is -- how valid can such an approach to evaluation be in providing ***accurate*** and ***useful*** information for decision making? This crucial question is taken up in two parts.

4.1 The Threat of Accuracy

Can outcome evaluation alone deliver on accuracy? Patton (1978, p. 149-50) provides what is perhaps an extreme answer. The case involves a demonstration program to teach welfare recipients the basic rudiments of parenting and household management. Following the adoption of this program by the state legislature, Patton noted that a highly respected independent research institute was contracted to evaluate the program. Researchers from this institute selected and interviewed a sample of the target population before and (18 months) after the program was adopted by the state legislature. The results obtained showed no measurable change and the legislature promptly terminated funding for the program. However, when Patton (1978, p. 150) delved further into the case, he found that:

> As it turned out, there is a very good reason why the program was ineffective. When the funds were initially allocated from the state to the city, the program became immediately embroiled in the politics of urban welfare. Welfare rights organizations

> questioned the right of government to tell poor people how to spend their money or rear their children.
>
> As a result of these and other political battles the program delayed and further delayed. Procrastination being the better of valor, the first parenting brochure was never printed, no household management films were ever shown, no workshops were held, and no case workers were ever trained. In short, *the program was never implemented -- but it was evaluated* (emphasis added).

Three threats to "construct validity" of outcome evaluation can be identified: (1) improper designation of intervention, (2) policy drift/transformation and (3) interaction of treatment and implementation results. The ways in which these threats act to prevent outcome evaluation from producing valid results are discussed below.

Improper Designation of Intervention. This is a threat "when a researcher designates an intervention to have occurred prior to implementation and then, analyzing outcome data, prematurely concludes that the policy is ineffective" (Musheno 1981, p. 79). It would seem obvious enough that evaluators would first confirm whether or not a program actually exists before evaluating it. But evidently this is not always the case as the observation by Patton was neither the first nor one of the very few. With examples from Title 1 of the U.S. Elementary and Secondary Education Act (ESEA), which had called for mandatory evaluation of its effects, McLaughlin's (1975) found that, in the vast majority of the evaluations, the "program" had not even taken place.

Knowing the specific point when the treatment or implementation occurs is one of the greatest threats facing evaluations that disregard the context of program implementation. In a world of incremental changes in which programs tend to be put into operation gradually, the problem presented by implementation lag or the time taken between policy adoption and actual program implementation is a major challenge for researchers of public programs. In their work, Cook and Campbell (1979, p. 227) suggest that a researcher should look for the *onset* of the treatment in designating intervention. But then what does "onset" mean? Does it simply mean the point at which the policy becomes formally adopted? In the case of the

Ontario seat belt law, for example, would intervention be designated at the date of adoption of the legislation by the Ontario law makers? Or would it be designated at a point much later, after there was adequate time for the law to be diffused to street-level bureaucrats (e.g., police officers) who would enforce it, and to the public who could not be expected to comply without knowledge of it?

Unfortunately, most applied researchers understand "onset" to mean either the point at which funding for the program is allocated or the date of its formal adoption by the authority responsible. Such a legalistic designation of intervention, as demonstrated by the case described by Patton, can be disastrous. Not only does it not recognize that program implementation is neither automatic nor certain, it also fails to realize that "causal thresholds can operate to make the manifestation of an effect dependent on reaching at least a certain level of the treatment" (Cook and Campbell 1979, p. 227). It fails to allow consideration for *widespread availability* of treatment to the target population, which is why "implementation, or assessing when the bureaucratic process has reached a point where target populations have a chance to be affected by the treatment, must be included in the designation of the intervention point for applied research" (Musheno 1981, p. 79). Only by doing this can applied researchers assure that implementation gap or delayed causation is taken into account in designating intervention.

Policy Drift/Transformation. This is a threat when "programs are so transformed in actual operation that to evaluate without finding out what actually transpires at the point of delivery may mean evaluating a program that is in fact very different from that originally intended" (Rossi and Wright 1977, p. 24). That the character of programs implemented nowadays is often changed in significant ways during implementation is not novel. In fact, with the advent of adaptive implementation (Berman 1980), the new conception of the implementation process as a "negotiating process" (Barrett and Fudge 1981, p. 20-6), "bureaucratic discretion at lower levels and pressures from powerful clients or constituency groups" (Sharp 1981, p.

102), it has come to be expected that programs should be revised and modified as they make their way through implementation. Hence, without understanding this variation in operational forms of programs, an evaluation can at best be an exercise in "abstract concepts" (Sorg 1981, p. 149).

In addition to the threat of programs being transformed when they undergo implementation, confusion over goals remains a significant threat. Although several suggestions have already been made as to how the problem should be dealt with, many of these centered on asking program staff to engage in a "goals clarification process." To the extent that "goals clarification" can become difficult and is considered to be "one of the most deadly ways to begin a program evaluation" (Patton 1982, p. 102), data on implementation can be relied upon to better discern program goals and character. The fact that program staff are likely to attempt to protect the organization, and that public programs are politically charged and often have undisclosed motives, suggest that behavioral data about implementation, on what was actually done as opposed to what was said to have been done (attitude), may in indeed be a preferable alternative for learning about program intents. The belief here is that, no matter how words are spoken, action speaks louder and, perhaps, more accurately.

Interaction of Outcome and Implementation Results. This is a threat when an evaluation "throws together results from sites where the program has been conscientiously installed with those from places which might have decided 'Let's not and say we did'" (Morris and Fitz-Gibbon 1978, p. 11). For any program evaluated and found ineffective without the knowledge of implementation, the crucial question remains -- Did the program and its theory fail or did its implementation? Applied researchers are now learning "from examining many social programs that the failure of a program may as often be due to faulty implementation as to ineffective treatment" (Rossi and Wright 1977, p. 23-4), which raises the question -- What should a program look like before it can be considered to be fully implemented (Patton 1978,

p.158)? That is, knowing that the implementation process always contains unknowns that change programs so that they look different from when initially conceived.

Evidently, in the parlance of causal language, only when it is known whether or not the "various practices and intervention efforts (are) undertaken as specified in the program design" (Bernstein and Freeman 1975, p. 19), can causal conclusions be made about a program's effectiveness that do not throw in results from faulty implementation. As neatly summarized by Shanks (1983, p. 44):

> In order to conclude either a failure of theory or program, the analysis considers the policy (transformed into program) as carried out. It is not possible to argue that a program did not activate a hypothesized causal process unless there is data on how the program was implemented. One must be sure that the program was actually implemented in accordance with the expectations of those hypothesizing the particular causal process.

4.2 The Threat of Non-utilization

Can outcome evaluation alone deliver on utility?[17] The answer is *yes*, but only for when decisions are about whether to continue or terminate a program, which is not most of the time.[18] Even when decisions are about that, they are much more likely to be made based on political fiats rather than evaluation results. So, when results from outcome evaluation could be used, they are not needed nor welcomed except for when they happen to agree with the political agenda. The following account by Goldenberg (1971, p. 388) is offered that answers

17. Of course there are a host of other reasons for non-utilization of evaluation results. For instance, no amount of quality information can bring about use if the evaluation client "had no interest in using evaluation results from the outset" (Weiss 1984, p. 159). "Ritualistic or "whitewash" evaluations are by no means uncommon with evaluation clients.

18. Enlarging a successful program would have been another use of outcome data but, as Patton (1978, p. 154)) points out, there are problems in this. He notes how a drug addiction treatment program, which had 100 per cent success rate for 2 years, was reduced to mediocrity after the budget of the program was tripled. Apparently, the basis of success for the program was its small staff and dedicated staff, which became lost once the program was enlarged.

the question about the utility of outcome evaluation results. The account describes one staff member's frustrations with the approach during a meeting:

> *Butch:* I've been sitting here and listening to you guys talk about research for about a half-hour now, and frankly the more I hear the more pissed-off I'm getting about all this research bullshit! Right now I don't give a good god-damn if I never hear another word about research or statistics for the rest of my life.... I've had it up to here with this research crap.... I am sick and tired of sitting here and being told that now, after being in the Center for about five months my kid is going to work 38 per cent more of the time than before he came into the RYC (Residential Youth Center). Big fucking deal! *What the hell does that tell me? ... what the hell does that tell people who don't know about the RYC, about the guts of this operation? Nothing. Nothing at all.* Its a lot of bullshit.
>
> *Kelly:* Butch, I think you are going overboard on this.
>
> *Butch:* The hell I am. I put my blood and guts into working with Ev and his crazy mother and all I come away with is that Ev is going to work 38 per cent more of the time. Big god-damn deal. *Where's Ev in all this? Where am I? Where's there anything that tells about what goes on between us day in day out?*
>
> *Kelly:* I think I know what's bugging you, Butch, but you know as well as I do that whether or not we're refunded depends a helluva lot on what these statistics show after a year.
>
> *Butch:* Look Scotty, you don't have to remind me about that. I know all about it. Washington wants statistics: CPI wants statistics; the whole world wants statistics. I know all that shit and I know its important. All I'm saying is that.... *if the research doesn't tell it like it is.... then from my point of view it isn't worth a shit* (emphasis added).

Apart from displaying what may be a typical attitude when an intended research consumer is at odds with the findings of research, Butch points to several threats to the use-value of outcome evaluation. These threats can be group and treated under *close vs distant causation relationships.*

Close vs Distant Causation Relationships. This is a threat when "evaluations can provide information indicating the extent of program success or failure, but not the reasons for that success or failure" (Thomas 1981, p. 96). In arguing "why look at program implementation," Morris and Fitz-Gibbon (1978, p. 10) assume a program to have worked and then ask the ensuing question: "What worked?"[19] In their opinion, if an evaluation could not answer this question, then it would have "wasted effort measuring the outcomes of an event which

19. Other related questions would be: How well did it work? Did all its variables work as well? Why did it work?

cannot be described and must therefore remain a mystery." An outcome evaluator can only respond that the program worked, no more no less. This is because of the problem that is being termed here as *distant causation*. It differs from *close causation* which, on the other hand, can be related to Weiss' (1984, p. 178-80) concept of "staying close to the scene." As Thomas (1981, p. 94) explains, close causation is so termed:

> .... because the cause-effect linkages being examined are relatively close to each other, in contrast to the frequent outcome evaluation focus on a single, relatively long cause-effect linkage of program inputs to outcomes. The idea is to examine each of the steps in the program's planned implementation and the activity levels at each of the program's major junctures. These examinations should, at a minimum, offer some clues on whether the program became enough of reality to have possibly produced any of the intended outcomes.

Thus it is that, without knowing whether or not intervention has occurred and, if it has, what brand of it occurred, and how well it occurred, an evaluator is likely to ask the wrong questions, select invalid effectiveness measures and, thereby, produce inaccurate results. Furthermore, without data on close causation relationships which could only be established through implementation evaluation, the results produced, even when accurate, are likely to be of limited use to decision makers in terms of enabling them to manipulate program variables to optimize policy goals. That is why, in order to avoid gathering useless and erroneous information, it is important to frame evaluation questions in the context of program implementation (Patton 1978, p. 175). The subsequent section of this paper discusses approaches to implementation evaluation and their methods of conduct, as a precursor of impact evaluation.

5. TYPES OF IMPLEMENTATION EVALUATION AND THEIR METHODS OF CONDUCT

In their path-breaking study on implementation, Pressman and Wildavsky (1973) noted that there had been minimal research carried out in the area. But since the time of this study, there has been an explosion of research on the subject, resulting in the proliferation, if not confusion, of terms for what can be included within the field (Thomas 1981, p. 92). However, for the intent of implementation evaluation, as the front-end part of outcome evaluation, it appears that the various terms discussed in the literature can be decomposed into four broad sequential types or categories. These are: (1) program documentation, (2) performance monitoring, (3) implementation assessment and (4) process evaluation. In the following pages, these forms of implementation evaluation and their methods of conduct are reviewed.

5.1 Program Documentation

Used by Morris and Fitz-Gibbon (1978, p. 11), the term deals with understanding the program and the environment in which it is operating. It basically covers descriptions for two topic areas of the program -- (1) contextual and (2) substantive features of the program -- and requires for the evaluator to collect and review all the printed material available on the program. The material to be reviewed may include relevant legislative acts, program plans or proposals, cabinet documents, policy papers, operations reports and internal memoranda, press releases, previous evaluation reports, annual reports, budget submissions, organization charts, relevant historical documents (i.e., academic works in the area) etc.. This beginning stage of the evaluation is to be approached in much the same way a thesis student would approach his work during literature search and review.

5.1.1 Contextual Features of the Program. These should address some of the *who, where* and *how* related questions about the program. While it is normally advisable, on the count of those not familiar with the program, for the context description to contain considerable amount of information, cursory attention can be afforded on some of the questions needed to establish program background and setting. The questions may include the following:

1. *Historical background and origins of the program*

* *When did the program originate? Did a similar program exist prior to this time? If the program is a modification of a previously existing program, what significant changes have been made and why?*

* *How did the program get started? Whose idea was it to implement the program? Was a formal or an informal needs assessment conducted and, if so, what were the results? What was the starting point for the needs assessment and who conducted it? Were members of the target population and local institutions involved in determining the needs or were the needs determined top-down?*

* *How was the program planned? Was there a formal planning process and, if so, who was charged with the responsibility of facilitating and coordinating this process? How participative was the process in terms of representing a board spectrum of interests and groups? What specific values or needs guided the direction of the planning process -- that is, what priorities were determined and how? What was the mechanism for translating broad statements of needs into tactical goals and objectives for the program?*

2. *The program's setting*

* *Where has the program been implemented? What sort of locales, e.g., state, region, communities or neighborhoods?*

* *What are the demographic characteristics of the affected locale -- for example, its population, density and growth rate, patterns of migration, ethnic composition, cohort groups (age and sex), etc.?*

* *What are the economic and social characteristics of the locale -- for example, the major occupations of people or economic base, unemployment rate and trend, size of average households and income, number of single parents, percentage of welfare recipients, number of schools and enrollments/withdrawals, literacy rate, social pathologies rates (e.g., crime, alcoholism, divorce, etc.), etc.?*

* *What does the political map of the locale look like? Does the locale have a structured form of government and, if so, how many such governments have jurisdiction within the locale? Are there any identifiable "attentive elites" that are known to influence policy matters in the locale?* (adapted and modified from Morris and Fitz-Gibbon 1978, p. 39-41).

5.1.2 Substantive features of the Program. These should describe what the program is supposed to do and the target population to which it is directed. Using the same sources of information as recommended previously, the evaluator should focus on questions such as follows:

1. Goals of the program[20]

• What is the legal basis for the policy in term of what it must, may and/or may not do?

• What was the program designed to accomplish (i.e., impacts and effects)? What operational objectives and targets were set, and under what time schedule?

2. The program's target group

• What population or segment of needs is to be served by the program? How homogeneous is this population and on what basis was it selected?

• Are members of the target population among the hard-to-study groups (e.g., poor)? In other words, what is the ease of collecting data on the population to demonstrate effect?

• Has the population previously been subject to a similar program or treatment? What other programs impact significantly on the population?

5.2 Performance Monitoring

This is the most quantitative of all the evaluation approaches. It normally refers to the "level" or "quantity" of activity (i.e., outputs) that takes place in a program but, used here, the term includes also the "quality" of that activity.[21] Therefore, rather than limit considerations to the physical units of goods/services produced and distributed under the program, the analysis is expanded to cover a complete inventory of the program in terms of "the amounts and kinds of program activities considered necessary for the accomplishment of program goals with a

---

20. See section 3.1 for additional questions on this.

21. The quality of performance *per se* is sometimes referred to as "effort" evaluation. Propounded by Suchman (1967, p. 61), the approach is supposed to represent an assessment of input or energy regardless of output.

particular stage of development" (Tripodi et al. 1971, p. 45). This is the question of determining "the degree to which inputs are available and operational at desired levels" (Patton 1978, p. 165).

> * *Program structure. What does the organizational chart convey about the staffing pattern of the program, particularly its flow of responsibilities? What offices and roles have been created? Are the links between the program elements fully explicated?*
>
> * *Capital Assets. What amounts of the office space, facilities and equipment devoted to the program are being made use of? Are the amounts devoted adequate?*
>
> * *Fiscal expenditures. How much of the funds allocated for both the program's capital and operating costs have been expended? How successful has the program implementors been in accessing additional sources of funding or resources?*
>
> * *Personnel. What kinds and numbers of personnel took part in the program? Were the numbers and qualifications of the personnel adequate for accomplishing the program's tasks? What special considerations were given in recruiting personnel for the program?*
>
> * *Activities. How much of the program elements or major work tasks/functions have been put into action? How much time (i.e., man-hours) and funds have been expended on responsibilities connected to these? What efforts have been made to solicit outside input on these from potential clientele, common public, experts and attentive publics, etc.?*
>
> * *Outputs. How much goods and services have been produced and distributed under the program? What can be guestimated about the quantity and quality of the outputs produced?* (adapted and modified from TBC 1983, p. 9; Morris and Fitz-Gibbon 1978, p. 41-3).

The obvious implication of monitoring performance is that a program is unlikely to have effects if it is relatively inactive, or if the right machinery for delivering it has not been installed.

5.3 Implementation Assessment

This is a term used by Williams (1976) to refer to the assessment of "the degree to which a field activity corresponds to its design specifications." The approach is sometimes referred to as *discrepancy evaluation* (Provus 1971). It compares the actual program with the ideal for the purpose of determining how far from the ideal the program can deviate and still be said to

have been implemented. Usually the focus is on what has been termed micro-implementation (Berman 1978), which describes the process by which lower level agencies or implementing bodies adapt, and adapt to, policies that they have been directed to execute.[22]

With data collected from "program documentation" serving as the "ideal" and data from "performance monitoring" as the "actual" program, the evaluator is finally ready to begin mapping the adaptation of policy as it progressed from broad ideas to concrete actions. The first step in this process is to discern critical disjunctures between the ideal and the actual program. After this is done, and if there exist significant divergences, the second step is to report the observation to the evaluation client who "may not want an evaluation to be conducted before the program is redirected to do what it is supposed to do, or may decide that an evaluation should still be conducted to determine the program impacts and effect" (TBC 1983, p. 10). The third and final step is to prepare to look for reasons that explain the critical disjunctures identified.[23] As noted by Thomas (1981, p. 92):

> To the extent that is possible, implementation assessment should go beyond these descriptive tasks to the analytical task of explaining any discrepancies between the expected and the actual program. Of particular interest is whether those discrepancies are based on (1) chance or accidental factors, (2) contextual factors (factors in the context of the local of implementation, or (3) factors integral to the program.

These considerations bring us to the final form of implementation evaluation, namely process.

---

22. The corresponding classification to micro-implementation is macro-implementation, which describes a different process involving "linked passages -- from policy decision to government programs, from program to local project adoption, from adoption to implemented local practice, and from practice to local outcomes" (Berman 1978, p. 167).

23. An alternate method for mapping adaptations of policies is that of Hall and Loucks (1978, p. 24) who suggest the following: (1) interviewing policy makers to determine their views of the essential components of the policy, (2) interviewing and observing a small number of implementors to discover the components of the policy as they have operationalized it, (3) from the two sets on interviews, preparing a formal interview guide to discover the variant components of the policy, and interviewing a large number of implementors, (4) constructing a checklist of policy components and completing it for each (or sample) implementor, and (5) analyzing the checklist results to determine the dominant configurations of the implemented policy (adapted by Sorg 1981, p. 147).

5.4 Process Evaluation

Used here, the approach focuses on the internal dynamics of and actual operations of the program in an attempt to explain changes in the program. Rather than look at the products, the approach concentrates on the process whereby the products are produced. As Patton (1978, p. 165) explains:

> Under field conditions in the real world, people and unforeseen circumstances shape programs and modify initial plans in ways that are rarely trivial. The process evaluator sets out to understand and document .... (and) to unravel what is actually happening in a program by searching for the major patterns and important nuances that give the program its character. A process evaluation requires sensitivity to both qualitative and quantitative changes in programs throughout their development ..... Process evaluations look not only at formal activities and anticipated outcomes, but also investigate informal patterns and unanticipated consequences in the full context of program implementation .... Finally, process evaluations usually include perceptions of people close to the program about how things are going. A variety of perspectives may be sought from people inside and outside the program.

Thus, for the first time in the evaluation, the evaluator is ready to change his strategy of research, from focusing on collecting and analyzing available documents on the program to entering physically and completely into the program's world in an attempt to explain why the variation in the program occurred in the first instance. To effectuate this strategy:

> the evaluator should meet with key people associated with the program: the directors, funders, related decision-makers, some clients and staff, other researchers, and other people with some connection to the program. Indeed, the most important aspect of this step is involvement with all types of program personnel. Moreover, the most important aspect of this involvement is listening to what these people say about the official goals of the program, the unofficial goals, the way the program is supposed to operate, the way it actually operates, and so forth (Conner 1984, p. 192-3).

This "listening and looking" phase is likely to require a significant amount of the evaluator's time but, once the phase is completed, the evaluator is finally ready to embark on an outcome evaluation (i.e., one that is bound to yield both accurate and useful results).

6. SUMMARY AND CONCLUSION

6.1 Summary

The overarching purpose of the paper was to synthesize a comprehensive approach to evaluating public policies that is not only capable of determining (un)success of programs, but also the explanations behind that (un)success. In establishing the background to this end, the field of public policy was defined as involving: (1) what the government chooses to do (outputs), (2) why it does it (intentions), (3) how it does it (process) and (4) what difference it makes (outcome). The origin of evaluation research was reviewed and found to have basis in the need for: (1) a more rational method of deciding which things amongst many are worth doing, and (2) a deeper understanding of those things (i.e., social problems) that form the objects of policy. The discussion of the role of social science suggested that the *raison d'etre* for evaluation research is to provide an objective judgment on the effectiveness of programs, so that necessary modifications could be made to improve performance.

In giving a step by step account of outcome evaluation, *the quasi-experimental interrupted time series design with non-equivalent groups* was identified as the recommended design under best experimental conditions. Otherwise, it was suggested that *the one-group pretest-posttest design with non-equivalent groups* should be used. Among the threats to validity that are normally considered in analyzing data from quasi-experiments (i.e., threats to internal, external, construct, statistical conclusion and conduct validities), an additional threat involving "approach validity" was identified. Under this new threat, several shortcomings to the outcome evaluation approach were noted including the possibilities for: (1) an improper designation of the intervention point, (2) a false understanding of the character of the implemented program, (3) the lumping together of faulty implementation with program results and (4) the inability to explain results obtained.

In proposing a solution to these threats, it was suggested that outcome evaluation should be framed in the context of program implementation. It was suggested that, unless it is known whether or not a program has been implemented, what mutated form of it was implemented, how well it was implemented in comparison to design specifications and what the dynamics (e.g., problems) of the program were, there may be little reason to expect an evaluation to yield accurate and useful results.

6.2 Conclusion

The basic premise of this paper was that there is a genuine need for evaluation as a means of determining program results and improving performance, and that this need has not been met because of the narrow focus on outcomes by past evaluations. Hence, the paper's resolve that developing a more comprehensive method of evaluation that meets high standards of research competence will better respond to the need. Unfortunately, however, "some investigators of research use have found scant relations between the quality of a study and its level of use" (Weiss 1984, p. 185), which leaves the contribution of this paper to be "improved quality and, consequently, potential use of evaluation research." In the words of Weiss (1984, p. 188), "that is no small contribution."

7. REFERENCES

Adie, R. & Thomas, P. (1982). ***Canadian Public Administration: Problematical Perspectives.*** Scarborough, ON: Prentice-Hall Canada Inc..

Barrett, S & Fudge, C. (eds.) (1981). ***Policy and Action: Essays on the Implementation of Public Policy***. New York: Methuen.

Berman, P. (1978). "The Study of Macro- and Micro-Implementation," ***Public Policy.*** Vol. 26:156-84.

Berman, P. (1980). "Thinking About Programmed and Adaptive Implementation: Matching Strategies to Situations," in Ingram, H. & Mann, D. (eds.) (1981). ***Why Policies Succeed or Fail.*** Beverly Hills, CA: Sage Publications Inc..

Bernstein, I. & Freeman, H. (1975). ***Academic and Entrepreneural Research.*** New York: Russell Sage.

Campbell, T. & Stanley, J. (1966). ***Experimental and Quasi-Experimental Designs for Research.*** Chicago: Rand McNally.

Chandler, M. & Chandler, W. (1979). ***Public Policy and Provincial Politics.*** Toronto: McGraw-Hill Ryerson Ltd..

Conner, R. (1984). "The Art of Implementation," in Rutman, L. (ed.) (1984). ***Evaluation Research Methods: A Basic Guide.*** Beverly Hills, CA: Sage Publications Inc..

Cook, T. & Campbell, D. (1979). ***Quasi-Experimentation: Design & Analysis Issues for Field Setting.*** Chicago: Rand McNally College Publishing Company.

Datta, L. (1976). "The Impact of the Westinghouse/Ohio Evaluation on the Development of Project Head Start," in Abt, C. (ed.) (1976). ***The Evaluation of Social Programs.*** Beverly Hills, CA: Sage Publications Inc..

Dimock, H. (1979). ***A Simplified Guide to Program Evaluation.*** Center for Human Relations and Community Studies, Concordia University, Montreal.

Dye, T. (1978). ***Understanding Public Policy***, 3rd ed.. Englewoods, N.J.: Prentice-Hall.

Evaluation Research Society (ERS), (1980). ***Standards for Evaluation*** (Draft). D.C.:ERS.

Floden, R. & Weiner, s. (1978). "Rationality to Ritual: The Multiple Roles of Evaluation in Governmental Processes," ***Policy Sciences.*** Vol. 9:9-18.

Goldenberg, I. (1971). ***Build Me a Mountain: Youth, Poverty and the Creation of New Settings.*** Cambridge, Mass.: M.I.T. Press. Cited in Durkin, R. & Durkin, A. (1975). "Evaluating Residential Treatment Programs for Disturbed Children," in Guttentag, M. & Struening, E. (1975). ***Handbook of Evaluation Research.*** Beverly, Hills, CA: Sage Publications Inc..

Grabe, S. (1983). ***Evaluation Manual.*** Paris: UNESCO.

Hall, G. & Loucks, S. (1978). ***Innovation Configuration: Analyzing the Adaptations of Innovations.*** Austin, Texas: R & D Center for Teacher Education, University of Texas.

Hogwood, B. & Gunn, L. (1984). ***Policy Analysis for the Real World.*** New York: Oxford University Press.

Jordon, J. & Sutherland, S. (1979). "Assessing the Results of Policy Expenditure: Program Evaluation in the Canadian Federal Government," ***Canadian Public Administration.*** Vol. 12(4):581-609.

Levy, F., Meltsner, A. & Wildavsky, A. (1974). ***Urban Outcomes.*** Berkeley, CA: University of California Press.

Mazmanian, D. & Sabatier, P. (1981). ***Effective Policy Implementation.*** Toronto: D.C. Heat and Company.

McCain, L. & McCleary, R. (1979). "The Statistical Analysis of the Simple Interrupted Time-Series Quasi-Experiment," in Cook, T. & Campbell, D. (1979). ***Experimental and Quasi-Experimental Designs for Research.*** Chicago: Rand McNally.

McLaughlin, M. (1975). ***Evaluation and Reform: The Elementary and Secondary Education Act of 1965/Title 1.*** Cambridge, Mass.: Ballinger.

Morris, L. & Fitz-Gibbon, C. (1978). ***How to Measure Program Implementation.*** Beverly Hills, CA: Sage Publications, Inc..

Musheno, M. (1981). " On the Hazards of Selecting Intervention Points: Time Series Analysis of Mandated Policies," in Palumbo, D. & Harder, M. (eds.) (1981) ***Implementing Public Policy.*** Toronto: D.C. Heat and Company.

Musto, L. (1985). "How to Use Interrupted Time Series to Assess the Effect of Ontario's Seat Belt Legislation on Road Fatalities," Mimeographed, Department of Political Science, Concordia University, Montreal.

Nachmias, D. (1979). ***Public Policy Evaluation: Approaches and Methods.*** New York: St. Martins Press.

Nadel, M. (1975). "The Hidden Dimension of Public Policy: Private Governments and the Policy-Making Process," ***Journal of Politics.*** Vol. 37:2-34.

Nakamura, R. & Smallwood, F. (1980). ***The Politics of Policy Implementation.*** New York: St. Martins Press.

Palumbo, D. & Harder, M. (eds.) (1981). ***Implementing Public Policy.*** Toronto: D.C. Heat and Company.

Patton, M. (1978). ***Utilization-Focused Evaluation.*** Beverly Hills, CA: Sage Publications Inc..

Patton, M. (1982). ***Practical Evaluation.*** Beverly Hills, CA: Sage Publications Inc..

Pressman, J. & Wildavsky, A. (1973). ***Implementation.*** Berkeley, CA: University of California Press.

Provus, M. (1971). ***Discrepancy Evaluation for Educational Program Improvement and Assessment.*** Berkeley, CA: McCutchan.

Reichardt, C. (1979). "The Statistical Analysis of Data from Nonequivalent Group Designs" in Cook, T. & Campbell, D. (1979). ***Experimental and Quasi-Experimental Designs for Research.*** Chicago: Rand McNally.

Rein, M. (1976). ***Social Science and Public Policy.*** New York: Penguin Books Ltd..

Rossi, P. & Wright, A. (1977). "Evaluation Research: An Assessment of Theory, Practice and Politics," ***Evaluation Quarterly.*** Vol. 1(1):5-52.

Rossi, P. (1972). "Testing for Success and Failure in Social Action" in Rossi, P. & Williams, W. (eds.) (1972). ***Evaluating Social Programs: Theory, Practice, and Politics.*** New York: Seminar Press.

Rossi, P., Freeman, H. & Wright, S. (1979). ***Evaluation: A Systematic Approach.*** Beverly Hills, CA: Sage Publications Inc..

Rossi, P., Wright, J. & Wright, S. (1978). "The Theory and Practice of Applied Social Research," ***Evaluation Quarterly.*** Vol. 2(2):171-91.

Rutman, L. (ed.) (1984). ***Evaluation Research Methods: A Basic Guide,*** 2nd ed.. Beverly Hills, CA: Sage Publications.

Scriven, M. (1971). "Pros and Cons About Goal-Free Evaluation," ***Evaluation Comment: The Journal of Educational Evaluation.*** Center for the Study of Evaluation, UCLA.

Shanks, G. (1983). ***Implementation of Public Policy.*** Unpublished PhD Dissertation, University of British Columbia.

Sharp, E. (1981). "Models of Implementation and Policy Evaluation: Choice and Its Implications," in Palumbo, D. & Harder, M. (eds.) (1981) ***Implementing Public Policy.*** Toronto: D.C. Heat and Company.

Smith, R. (1981). "Implementing the Results of Evaluation Studies," in Barrett, S. & Fudge, C. (eds.) (1981). ***Policy and Action: Essays on the Implementation of Public Policy***. New York: Methuen.

Sorg, J. (1981). "Pursuing Policy Optimization by Evaluating Implementation: Notes on the State of the Art," in Palumbo, D. & Harder, M. (eds.) (1981) ***Implementing Public Policy.*** Toronto: D.C. Heat and Company.

Statistics Canada. (Annual 1970-83). "Causes of Death," Catalogue 84-203.

Suchman, E. (1967). ***Evaluation Research: Principles and Practice in Public Service and Social Action Programs.*** New York: Russell Sage.

The Toronto Star. (Aug. 18, 1977). "Seatbelts Keep Saving Lives," p. A14.

Thomas, J. (1981). "'Patching Up' Evaluation Designs: The Case for Process Evaluation," in Palumbo, D. & Harder, M. (eds.) (1981) ***Implementing Public Policy.*** Toronto: D.C. Heat and Company.

Tien, J. (1979). "Towards a Systematic Approach to Program Evaluation Design," ***IEEE Transactions on Systems, Man and Cybernetics.*** Vol. 9(9):494-515.

Treasury Board of Canada (TBC), (1983). ***Principles for the Evaluation of Programs.*** Government of Canada, Ministry of Supply and Services.

Tripodi, T., Phillip, F. & Irwin, E. (1971). ***Social Program Evaluation: Guidelines for Health, Education, and Welfare Administration.*** Itasca, Ill.: F.E Peacock.

Weiss, C. & Rein, M. (1970). "The Evaluation of Broad-aim Programs: Experimental Design, Its Difficulties, and an Alternative," ***Administrative Science Quarterly.*** Vol. 15:97-109.

Weiss, C. (1972). ***Evaluation Research: Methods of Assessing Program Effectiveness.*** Englewoods, N.J.: Prentice Hall.

Weiss, C. (1984). "Increasing the Likelihood of Influencing Decisions," in Rutman, L. (ed.) (1984). ***Evaluation Research Methods: A Basic Guide.*** Beverly Hills, CA: Sage Publications Inc..

Weiss, C. (ed.) (1977). ***Using Social Research in Public Policy Making.*** Lexington, Mass.: Lexington Books/D.C. . Ieath.

Wildavsky, A. (1979). ***Speaking Truth to Power: The Art and Craft of Policy Analysis.*** Toronto: Little, Brown and Company.

Williams, W. (1976). "Implementation Analysis and Assessment," in Williams, W. & Elmore, R. (eds.) (1976). ***Social Program Implementation.*** New York: Academy Press.

Worthen, B. & Sanders, J. (eds.) (1973). ***Educational Evaluation: Theory and Practice.*** Ohio: Charles A. Jones.